DOCUMENT RESUME

ED 152 853 TM 007 113

AUTHOR Gavin, Anne T.
TITLE Guide to the Development of Written Tests for Selection and Promotion: The Content Validity Model. Technical Memorandum 77-6.
INSTITUTION Civil Service Commission, Washington, D.C. Personnel Measurement Research and Development Center.
PUB DATE Jun 77
NOTE 44p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Ability Identification; Equivalency Tests; Government Employees; *Job Analysis; *Job Skills; *Occupational Tests; *Personnel Selection; Pretesting; Promotion (Occupational); Scoring; Skill Analysis; *Test Construction; Testing; Test Interpretation; Test Reliability; *Test Validity; Vocational Aptitude
IDENTIFIERS *Content Validity

ABSTRACT

Content validity involves a judgment that all knowledges, skills, abilities, and other worker characteristics (KSAOs) needed for successful job performance (or a specified portion, the job content domain), have been adequately sampled by the tests used for employee selection and promotion decision making. The model outlines the procedure for documenting the content validity of both an individual test per se and the total employee selection program. Major steps are: (1) Job Analysis--identifying KSAOs; determining their importance, level, and ability to differentiate among workers; (2) Selection Plan--determining the appropriate measurement criteria, for example: written test, performance test, physical requirements, interviews, evaluations of background and experience, then combining measurement results; (3) Test Plan--defining the job content domain, test objectives, format, timing, length, and sampling method; (4) Test Construction--item writing and review, test assembly and pretesting; (5) Test Use--administration and scoring, determining technical adequacy and periodically reviewing for currency. A representative panel of subject matter experts must be involved in each step. This model meets content validity standards set by the American Psychological Association and federal regulatory agencies. Bibliographical references, a brief glossary, and sample forms are appended. (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Personnel Research and Development Center

Technical Memorandum 77-6

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED152853

TM007 11

Guide to the Development of Written Tests for Selection and Promotion: The Content Validity Model

United States
Civil Service Commission
Bureau of Policies and Standards

Technical Memorandum 77 - 6

Guide to the Development of Written Tests for Selection and Promotion: The Content Validity Model

Anne T. Gavin

Personnel Research and Development Center
United States Civil Service Commission
Washington, D. C. 20415

June 1977

Guide to the Development of Written Tests
for Selection and Promotion: The Content
Validity Model

ABSTRACT

This report describes a procedure which examiners in State and local governments may wish to follow in using test materials provided by the U. S. Civil Service Commission and from other sources in the development of content valid and job-related written employment tests.

Preface

This manual presents a model for written test development based on the requirements for demonstrating content validity. It is intended for use by examiners in State and local government personnel departments who develop and use written tests for selection and promotion. The procedures preceding and encompassing written test development are outlined as explicitly as possible, keeping technical presentations to a minimum; it is assumed, however, that the reader has some familiarity with personnel measurement and test development. Depending on the training and experience of the reader, some steps will be more familiar than others; however, all of the steps included are necessary for proper test development. It is recommended that personnel responsible for test development have a working knowledge of such fields as elementary psychology, basic principles of testing, individual differences, industrial psychology, elementary statistics, and basic test theory. Generally, such preparation would require at least a master's degree in psychology or education, with significant course work and experience equivalent to a major in educational and psychological measurement. Since the professional soundness and legal defensibility of written tests often depends on the qualifications of those making decisions in the test development process, it is suggested that readers seek the services of an outside measurement consultant for those steps beyond their expertise and the personnel and technical resources of their jurisdictions.

The sections of this manual detailing test construction procedures are based in part on guidance materials prepared by Sandra Payne (1976) for applying the content validity model in developing promotion tests for firefighters. Additional assistance in the preparation of this manual was provided by members of the staff of the Personnel Research and Development Center and the Regional Offices (particularly the regional psychologists) of the U. S. Civil Service Commission who carefully reviewed an earlier draft and and offered helpful suggestions for revisions.

The author wishes to thank those colleagues and other readers of the earlier draft who made many constructive suggestions, particularly Dr. Albert Maslow of the Educational Testing Service. Although such contributions undoubtedly improved the final report, responsibility for the contents rests solely with the author.

TABLE OF CONTENTS

	Page
Introduction	1
Concepts of Validity and Job Relatedness	1
Definitions of Test Validity	2
Job Relatedness and Content Validity	2
The Content Validity Model	3
Application of the Content Validity Model in Written Test Development	5
Identification of the Relevant Content Domain(s) of the Job	7
Job Analysis	7
Identification of Knowledges, Skills, and Abilities for Measurement	8
KSAO coverage	11
Determining importance of KSAOs	11
Determining level of KSAOs	11
Determining KSAOs which differentiate	11
Development of a Selection Plan	12
Determination of Appropriate Measurement Instruments	12
Combining Measurement Results	14
Preparation for Test Development	15
Size of Panel	15
Qualifications of Panel Members	16
Documentation File	16
Training	16
Security	16
Development of the Test Content Outline	17
Definition of the Job Content Domain	17
Definition of Test Objectives	17
Planning the Format, Timing and Length of the Test	18
Format of test questions	18
Timing of test	18
Length of test	18
Planning the Content Area Coverage of the Test	19
The Test Content Outline	20
Test Construction	20
Item Writing	20
Criteria for Item Review	20
Test Assembly	21
Pretesting	22

TABLE OF CONTENTS

	Page
Test Administration and Scoring	23
Test Administration	23
Scoring	23
Setting Cut Scores	23
Technical Adequacy of Test	24
Frequency Distribution	25
Measures of Central Tendency	25
Measures of Variability	25
Measures of Reliability	26
Item Analysis	26
Review of Selection Program	26
Glossary	28
References	31
Bibliography	32
Appendix A: Sample Selection Plan Format	34
Appendix B: Summary of Documentation File Contents	35
Appendix C: Sample Security Form	36
Appendix D: Sample Test Content Outline	37

Guide to the Development of Written Tests for Selection and Promotion: The Content Validity Model

This report describes a model for written test development which can be used to support a claim of content validity. The first part of the report discusses the theoretical issues pertaining to validity, job relatedness and the rationale underlying the content validity model. A detailed discussion of these issues was considered necessary in light of the usually inadequate and simplistic treatment they have received in the past, leading to confusion about the appropriateness and proper application of the content validation strategy in employment situations. Later sections present the requirements for documenting content validity, starting with the decision to use a written test and other preliminary determinations. The application of the model in the actual test development process is treated in detail.

The content validity model consists of a sequence of steps to be followed in the development of a selection program which includes a written test in order to link the test to the job. The presentation emphasizes the information provided by each of these steps to document the validity and job relatedness of the test. The specific procedures through which this information is actually obtained, such as job analysis, specification of the knowledges, skills, abilities and other worker characteristics (KSAOs) to be measured, and the selection of appropriate measurement instruments are discussed more briefly. The reason for this emphasis is twofold: (1) There are often a variety of acceptable procedures available for securing this information and the choice of one for presentation could be considered arbitrary and perhaps misleading; and, (2) The procedures are often complex, would require extensive training materials for complete explication, and most likely are already familiar to test developers.

The test development model presented in this paper meets the requirements of major test guidelines for demonstrating content validity for written tests. The American Psychological Association's (APA) Standards for Educational and Psychological Tests (1974) provide a professional guide to good testing practices for both test developers and users, and the Principles for the Validation and Use of Personnel Selection Procedures (1975) published by the Industrial-Organizational Division of the APA further elaborate these requirements in the employment situation. Guidelines on testing standards have been issued by the Federal regulatory agencies and are based on these APA publications: The Federal Executive Agency Guidelines on Employee Selection Procedures, which have been adopted by the U. S. Department of Justice, the Office of Federal Contract Compliance, and the U. S. Civil Service Commission, and the Guidelines on Employee Selection Procedures, which are issued by the Equal Employment Opportunity Commission.

Concepts of Validity and Job Relatedness

Selection involves making a decision about whom to hire and whom to reject for a particular job. In this report the term "selection" is used in reference to both entrance and promotion decisions. Selection is made on the basis of available information which indicates the likelihood that an applicant will be able to perform the job, and in many cases, the relative standing (rank) of applicants in terms of this potential. In order to maximize the precision and job relatedness of selection decisions, employers should appraise as many applicant qualifications critical to successful performance of the various aspects of the job as is feasible and practical. (Aspects of the job include such things as tasks, duties, physical requirements, work environment, and required knowledges, skills, and abilities.) The validity of the total selection program depends on the validity of each of the individual assessment instruments (tests) used to measure applicants in the various aspects of job performance. Although the validity of any one of these components may not be sufficient to insure the validity of the total program, tests which are not valid can only reduce the accuracy and utility of selection decisions. Therefore, every effort should be made to validate tests used in employment

situations and to use only tests of proven (or provable) validity in making selection decisions.

Definitions of Test Validity

There are three generally recognized types of validity: content, construct and criterion-related. Each type of validity refers to the appropriateness of a particular type of inference which is drawn from test scores (American Psychological Association, 1974, p. 25). Content validity applies to inferences made on the basis of test scores regarding present level of achievement in terms of knowledges, skills or abilities necessary for successful job performance and is based on the correspondence between job content and test content. Construct validity involves inferences that particular constructs determined to underlie job performance are also measured in test performance and therefore, that the test scores describe such aspects of the persons measured. (A psychological construct is a theoretical idea or concept used to explain and/or organize some aspect of existing knowledge; more than the label, such as "anxiety" or "clerical aptitude," the construct is a dimension understood or inferred from its network of interrelationships with other measures and must be precisely defined.) Criterion-related validity applies to inferences made from test scores regarding a person's estimated or predicted performance on a different, operationally independent aspect of a job (criterion). Strategies used to demonstrate each of these types of validity involve judgment. Even those strategies resulting in quantitative "validity coefficients" involve judgment in evaluating the appropriate criteria for study and the resulting estimates of validity.

It is not correct to speak of the validity of a test. Test validity is dependent upon the type of test, the manner in which the test is used and combined with other assessment results, and the type of inference drawn from the test scores. The strategy used to demonstrate a test's validity should be appropriate to the inferences drawn from that test's scores. For example, where test scores are used to indicate present level of proficiency in the knowledges, skills or abilities needed to perform a job, validity is appropriately demonstrated through content validation. In planning a selection program, it is important to remember that the way in which test scores are to be used often dictates the type of validation strategy necessary to support the program.

In general, the closer the behaviors needed in test performance match those needed in job performance, the stronger the evidence of content validity. In cases where this connection is not direct and therefore, where scores on the test cannot be used to infer present level of achievement, other validation strategies are needed. For example, if a test of general reasoning ability is used to measure how well a candidate will learn to perform certain job tasks involving reasoning, it must be shown that scores on the test are empirically related to performance on the task (criterion-related validity) or that performance of the job tasks requires the same general reasoning ability as does test performance (construct validity). On the other hand, if a representative sample of the actual job tasks were administered to candidates, a content validation strategy would be appropriate.

Job Relatedness and Content Validity

According to the classic concept of content validity, a conclusion of validity is based on the way in which a test is developed and the extent to which it constitutes a sample of a specified domain. For example, a written test which contains items which adequately sample a specified knowledge area would be considered a content valid measure of that knowledge area in the classic sense of content validity. However, unless the knowledge is used in a job and is necessary for successful performance, such a content valid test will not be job related. The concept of job relatedness means that a test used in making a selection decision must measure knowledges, skills, abilities or other worker characteristics needed for successful job performance. The requirement for job relatedness means that tests which provide valid information about applicants can be used in employment situations only if that information is relevant to job performance. Extending the classic concept of content validity for employment purposes therefore requires matching job require-

ments with test content. In this paper the use of the term content validity will be in the extended sense, i.e., that it includes job relatedness; therefore, when a content validation strategy is used, no separate documentation of job relatedness is necessary since a content valid test is by definition job related. When other validation strategies are used to support the use of a test, however, the content validity model may also be useful as a method of documenting job relatedness since it provides a means of linking the job content to the test content.

The Content Validity Model

Content validity involves judgment that the job content universe or a specified job content domain has been adequately sampled by the measurement instrument(s) used in making employment decisions. The job content universe consists of "the total job; everything, known and unknown, which the incumbent does and must know in order to do it," (American Psychological Association, Division of Industrial-Organizational Psychology, 1975, p. 17). It is generally neither feasible nor practical to consider the entire job content universe in making personnel decisons. Rarely are all aspects of the job known, and often there are aspects of job performance which are impossible to measure or which yield little information helpful in making selection decisions. For these reasons, content valid tests generally sample definable, non-trivial subsets of the job content universe called "job content domains."

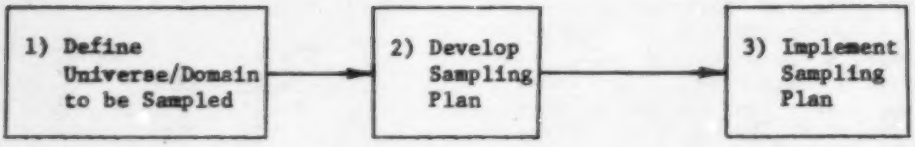
Job content domains may be defined broadly or narrowly depending on the purpose of the definition. Some examples of job content domains are: all job tasks; critical job tasks; knowledges needed to perform the job; work products; physical requirements of job performance; or a functional subgrouping of job tasks (such as "tasks relating to test development" for a personnel specialist). Certain types of measurement instruments designed to be content valid are appropriately used to sample only certain job content domains. For example, performance tests generally sample the domains of job tasks or of skills needed for successful job performance; strength and stamina tests and medical examinations sample the domain of physical requirements for job performance; written tests are most appropriate for sampling the domains of knowledges and skills needed in job performance.

There is an important distinction (Federal Executive Agency Guidelines on Employee Selection Procedures, Sec. 12c(3)), between the content validity of an individual test per se (based on the adequacy of sampling from a specified job content domain) and the content validity of a total selection program (which may consist of a single test) for a particular job (based on the degree to which the job domain(s) sampled represent critical behaviors or a substantial portion of the job). Since written tests are usually content valid for the knowledge or skill job content domains only, if a selection program consists solely of a content valid written test it must be shown that the domain it samples is critical to job performance or represents a substantial proportion of the job content universe. Where such a showing cannot be made, use of the test as part of a multipart selection program which fully assesses critical aspects of job performance or a greater proportion of the job is indicated.

To the degree that a test used as part of a multipart selection program is demonstrated to be content valid for a particular domain, it contributes to the validity of employment decisions based on such a program. To the degree that any assessment instrument used in a selection program is shown to be related to job performance, the job relatedness of the program is enhanced. The content validity model presented in this paper outlines procedures for providing the information to make both these determinations. The procedures used in the development of the selection plan allow for documentation of the job relatedness of the selection program and for the specification of domains critical to job performance or representing a substantial proportion of the job content universe; the procedures used in the development of the written test provide the documentation of the content validity of that test for a particular job-related domain.

The content validity model consists essentially of three steps into which the procedures for the development of both the selection plan and the written test fall:

STEPS



PROCEDURES

Selection Plan:	. Job analysis . Identify knowledges, skills, and abilities for measurement	. Specify domain(s) to be sampled . Specify measurement objectives . Specify measurement instrument(s) to be used to sample domain(s) . Specify strategy for combining measurement results	. Establish minimum education and/or experience qualifications . Develop assessment instruments (written test, rating guide, etc.)
Written Test:	. Define job content domain to be measured . Specify test objectives (from selection plan)	. Assign proportional number of items (weight to be assigned) to each area of job content domain . Decide on appropriate item type(s), length, timing and format of test	. Write/review items . Assemble test . Check technical adequacy of test

The establishment of content validity is essentially an inference, based on the information available from the steps outlined above, of the adequacy of the sampling process. In the case of establishing content validity for an entire selection program, a judgment is made regarding the degree to which critical and/or substantial portions of the job are covered by the measurement instrument(s) used. For written tests, the inference of content validity requires a judgment that the specified job content domain has been adequately sampled by the test. The issue is one of reasonable (not statistical) representativeness.

"Representative" refers to both the types of behaviors assessed and the proportional coverage of different elements (KSAOs) in the sampling plan. By definition, content valid tests require competitors to exhibit behaviors needed for job performance. In a performance or job sample test, for example, the test situation may virtually duplicate the job situation and hence, require identical behaviors. The written test situation, on the other hand, necessarily requires some behaviors which may not be required on the job; for example, behaviors needed for marking answers. Insofar as test performance requires the same knowledges and skills needed for job performance, however, the test content domain may be considered a sample of the job content domain. The degree to which such a sample is judged representative of job behaviors depends on

the judgment of experts and requires an unambiguous definition of the job content domain in terms of behaviors with generally accepted meaning.

The representativeness of sampling also refers to the degree to which at least the critical and/or substantial aspects of the job or specific job domain are measured in proper proportion. In order to judge the adequacy of sampling in this regard it is necessary to determine the importance of the job content domains and of the elements (specific knowledges, skills, or abilities) within each domain that is sampled. This information is needed in order to judge the degree to which the coverage provided by the total selection program and/or the written test is representative of the total job or particular job content domain in terms of critical or substantial portions of each.

At the present time, the essence of establishing content validity is primarly judgment. While methods for quantifying the degree to which the test content domain is a representative sample of the job content domain have been suggested, such quantification merely formalizes judgment into the degree of agreement among qualified judges. The qualifications necessary in order to make judgments of content validity primarily relate to knowledge of the job content domain. Therefore, qualified judges are usually those who have performed the job, supervised it, or have carefully analyzed relevant parts of it.

The final judgment of content validity rests on evidence pertaining to the procedures followed in the development of the selection program and individual assessment instruments. Although this judgment cannot be made until the sampling plan has been implemented, the establishment of content validity involves all the processes related to test development. The establishment of content validity for a test therefore rests on a showing that it was developed according to procedures similar to those outlined in this manual; without the information these procedures provide, a judgment of content validity usually cannot be made or supported. The development of the test should be planned and carried out with the goal of establishing content validity. The important point is that it is usually impossible to establish content validity for a test which lacks adequate documentation of the procedures by which it was developed.

Application of the Content Validity Model in Written Test Development

The content validity of a written test is therefore based on an evaluation of the steps taken in the test development process to ensure that the job content domain has been adequately defined (job analysis) and sampled (test development). Simply stating that the test was developed according to a model for content validity is not sufficient. Documentation of the decisions made throughout the process, the definition of the job content universe, how the test content relates to the job content domains, the choice of individual test items, and the professional competence of the persons making these decisions must be provided in support of the statement. A summary of the steps to be followed in developing written tests using the content validity model is given in Table 1. The remainder of this report focusses on the application of the content validity model to the development of written tests.

There are decisions which must be made in the development of content valid tests prior to the ultimate inference of validity. The major decision points are the specification of the job requirements in the job analysis phase, the specification of the measurement requirements (linking the job requirements to the assessment instruments) in the selection plan, and the specification of the written test requirements (determining the adequacy of the sampling) in the test outline. These decisions may appropriately be made by panels of subject matter experts (SMEs). In most cases the qualifications needed by panel members relate to extensive knowledge of the job content domain(s). Very often the same panel makes all these decisions in the test development process (although additional members with specialized expertise may be needed at some stages), including

Table 1

Summary of the Content Validity Model for Written Test Development

I. Job Analysis

A. Identify knowledges, skills, abilities and other worker characteristics needed for successful job performance.
B. Specify important and/or critical KSAOs, the level of each, and indicate which KSAOs differentiate between better and poorer job performance.
C. Develop list of behaviorally defined KSAOs to be assessed in selection program.

II. Selection Plan

A. Specify measurement instruments to be used.
B. Determine knowledges and/or skills to be measured by written test.
C. Decide how to combine results of measurement instruments (if using multipart assessment program).

III. Test Plan

A. Define job content domain to be sampled by test.
B. Define test objectives.
C. Decide on format, length, and timing of test.
D. Specify method of sampling from the job content domain for test.
E. Document the test content outline specifying the correspondence between test content and job content.

IV. Test Construction

A. Develop pools of test items to meet specifications of test content outline.
B. Carefully review all items.
C. Choose items for inclusion in test.
D. Assemble test and prepare official answer key.

V. Test Use

A. Provide for standard administration and scoring.
B. Determine technical adequacy of test.
C. Periodically review currency of test.

the final inference of content validity. Therefore, the qualifications of panel members for each decision point are not detailed in this report, except for the actual test assembly phase. It should be emphasized, however, that at any stage in which a panel is used, members must be carefully selected on the basis of their qualifications and representativeness, and the basis for their selection fully documented. Where panels are used to make decisions in the test development process, the extent of agreement among panel members should also be reported in the documentation records.

Identification of the Relevant Content Domain(s) of the Job

Essential to any claims of validity, regardless of the validation strategy involved, is evidence that the selection program adequately covers important aspects of the job. This documentation is provided in the selection plan. The development of the selection plan begins with the determination of the aspects of the job which must be measured to select the best applicants. In terms of the content validity model, this involves the definition of the job content domain(s) to be sampled by the measurement instruments used. There are many job content domains which are appropriately measured by means other than written tests. Such aspects of the job should be covered in the selection plan if they are critical or constitute substantial portions of the job, although they may be defined in ways different from those discussed in this paper.[1]

Job Analysis

The definition of job content domain(s) measured by written tests begins with a job analysis. A properly conducted job analysis is necessary in order to:

1) Establish the degree to which the job for which people are to be selected is homogeneous across different organizational and physical locations; that is, establish that the "job" does not actually consist of relatively unique positions under the same label, but that there is a content universe common to all individual positions of the same job.

2) Specify as completely as possible the content of the job.

3) Define the boundaries and content of each job content domain analyzed.

It is important that the job analysis method chosen provide information for defining the job content domains specifically for selection or promotion. Analyses done for pay or classification purposes may not be relevant to selection issues, and job descriptions generally contain only very limited information about characteristics of successful job performance.

There are a number of professionally accepted techniques for conducting a job analysis for selection. A discussion of the relevant advantages and disadvantages of each is beyond the scope of this report. Good sources of information on job analysis techniques can be found in standard textbooks in industrial psychology. An article by Prien and Ronan (1971) reviews many of these procedures and the chapter on job analysis in Wright (1974, Chap. 3) offers a detailed discussion of this topic.

The job analysis technique chosen for use must be appropriate for the development of content valid tests. That is, it must provide valid and reliable information about the job content universe and specific job content domains in an objective manner. The reliability and validity of job analysis information depends in large part on the meeting of the following criteria:

1) use of a systematic and appropriate data collection method

---

[1]Note that content validity is not restricted to written tests, but may be the appropriate strategy for establishing the validity of any assessment instrument which samples job content, such as rating of education and experience or a performance test.

2) collection of data from an adequate sample of incumbents or others knowledgeable about the job

3) verification of the currency, accuracy, and representativeness of data

4) concurrence of more than one qualified analyst on findings

The meeting of these criteria should be carefully and completely documented. The job analysis essentially provides an operational definition of the job content domain(s). The documentation record of the process by which this definition was derived should include at a minimum:

- o job title
- o performance (grade) level(s)[1]
- o date of job analysis
- o names and qualifications of analysts
- o complete description of job analysis technique(s) and instruments used
- o source of all information collected about job
  - documents used (e.g., position descriptions, training manuals, work samples)
  - identification of job incumbents and supervisors contributing information, including job title and level, unit, qualifications and total sample size
- o results of job analysis--may be listing of duties or KSAOs (Where KSAOs have been inferred from job duties or tasks this process and the linkage between duties or tasks and KSAOs must be carefully documented. It is always preferable to list the job duties. This will allow for monitoring the job over time and to determine if changes are later needed in the selection plan.)

Identification of Knowledges, Skills, and Abilities for Measurement

The content validity model requires that the definition of the job content domain(s) for measurement be derived from observable aspects of actual job performance. For purposes of measurement, job content domains are generally defined in terms of the knowledges, skills, abilities, and other worker characteristics (KSAOs) needed to successfully perform the job. These KSAOs must be linked to job behaviors, observable aspects of job performance. Information from the job analysis is used to provide this linkage. There are two methods used to derive KSAO information from job analysis data: direct identification of necessary KSAOs in the job analysis process or inference of necessary KSAOs from job analysis data regarding duties, tasks, or responsibilities. The method by which the KSAOs are derived may affect the choice and definition of job content domains for measurement as shown in Figure 1 and discussed below.

A job content domain has been defined as some subset of the job content universe which has specified boundaries so that its contents can be unambiguously defined.

---

[1]In rare cases the job analysis may also include an examination of job levels to which applicants might normally be expected to progress within 5 years. (See the Federal Executive Agency Guidelines on Employee Selection Procedures, Sec. 5g for further information on this point.) It may be appropriate to test in the initial selection process for aspects of the higher level job which are not needed or taught in the lower levels. (This is most commonly the case in fairly complex jobs requiring knowledges which the candidate must possess upon entry, but which will not be used until performing at a higher job level.) For content validity purposes it should be shown that such higher level requirements must be possessed upon entry at the lower level (i.e., they normally would not be acquired either on or off the job between entry and advancement to higher level). Such determinations are often difficult to verify and document, and therefore should be made with extreme care.

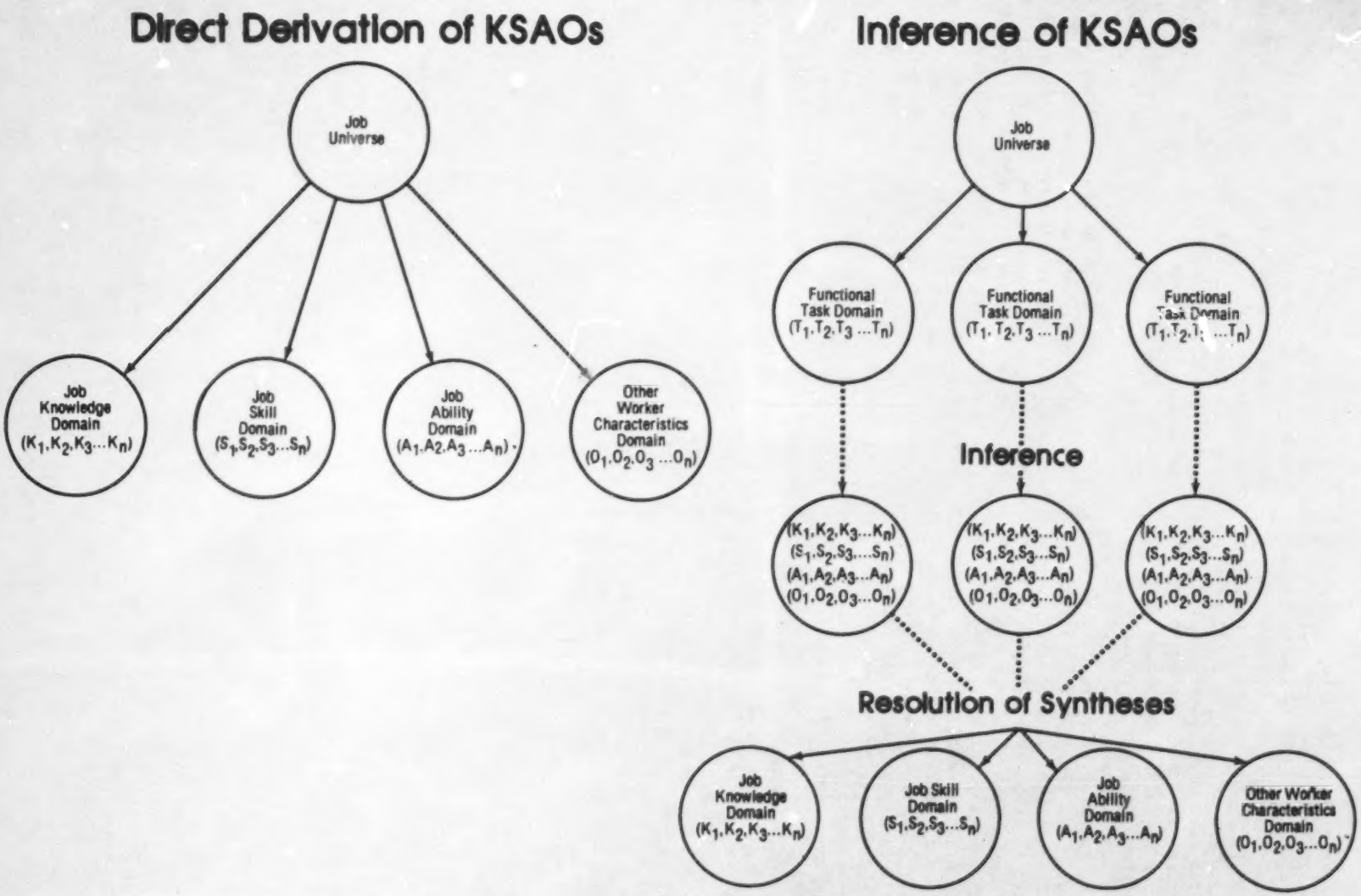

Direct Derivation of KSAOs
Job Universe
Job Knowledge Domain ($K_1, K_2, K_3 \ldots K_n$)
Job Skill Domain ($S_1, S_2, S_3 \ldots S_n$)
Job Ability Domain ($A_1, A_2, A_3 \ldots A_n$)
Other Worker Characteristics Domain ($O_1, O_2, O_3 \ldots O_n$)
Inference of KSAOs
Job Universe
Functional Task Domain ($T_1, T_2, T_3 \ldots T_n$)
Functional Task Domain ($T_1, T_2, T_3 \ldots T_n$)
Functional Task Domain ($T_1, T_2, T_3 \ldots T_n$)
Inference
($K_1, K_2, K_3 \ldots K_n$) ($S_1, S_2, S_3 \ldots S_n$) ($A_1, A_2, A_3 \ldots A_n$) ($O_1, O_2, O_3 \ldots O_n$)
($K_1, K_2, K_3 \ldots K_n$) ($S_1, S_2, S_3 \ldots S_n$) ($A_1, A_2, A_3 \ldots A_n$) ($O_1, O_2, O_3 \ldots O_n$)
($K_1, K_2, K_3 \ldots K_n$) ($S_1, S_2, S_3 \ldots S_n$) ($A_1, A_2, A_3 \ldots A_n$) ($O_1, O_2, O_3 \ldots O_n$)
Resolution of Syntheses
Job Knowledge Domain ($K_1, K_2, K_3 \ldots K_n$)
Job Skill Domain ($S_1, S_2, S_3 \ldots S_n$)
Job Ability Domain ($A_1, A_2, A_3 \ldots A_n$)
Other Worker Characteristics Domain ($O_1, O_2, O_3 \ldots O_n$)

Figure 1. This figure shows two ways of linking job content domains defined in terms of knowledges, skills, abilities, and other worker characteristics to the job content universe.

When KSAOs are directly derived in the job analysis stage (by use of such techniques as the job element method or the Position Analysis Questionnaire, for example), the domains tend to be content-bound, e.g., the job knowledge domain, the job skill domain, etc. Written test development usually requires sampling from such content-bound domains. Use of duty or task based job analysis methods requires an additional step in the identification of KSAOs, since they must be inferred from the duty or task information. This linkage or "inferential leap" from observable job tasks, behaviors, duties, etc. to unobservable KSAOs assumed to underlie such behaviors is a critical point in the development of a selection program and written test. There is no single method for making and documenting the accuracy of these linkages. Since this is essentially a judgmental process, however, it is generally recommended that a number of persons familiar with the job be called upon to infer KSAO requirements in some kind of rating procedure. Such rating procedures are often incorporated into the job analysis to insure the raters' familiarity with the job and to avoid the necessity of convening an additional panel. Because this step provides the essential link between the job content domain(s) and observable aspects of job performance, the criteria for deriving the KSAOs and the qualifications of those making such determinations must be carefully documented.

As noted above, such rating procedures usually provide a direct linkage of KSAOs to job tasks or duties. If the tasks had been grouped before rating according to such criteria as functional similarity or criticality, job content domains may be task-bound, e.g., KSAOs needed for equipment operation, KSAOs needed for equipment repair, etc. The content validity model does not preclude the redefinition of such content domains as long as the boundaries and content of each domain are specified. Therefore, the contents of the task-bound domains cited above could be synthesized and reallocated into content-bound domains which underlie tasks across all functional groupings; e.g., knowledges needed in job performance, skills needed in job performance, etc. The only restraint is that the final set of domains be as mutually exclusive as possible. Since written tests are appropriately used to sample primarily knowledges and skills (and in some cases, abilities as noted below), when job content domains are being defined for such measurement purposes, they should usually be defined in terms of knowledges and skills (worker qualifications) rather than in terms of tasks or duties (job requirements).

It should be noted that inferences about necessary aptitudes, personality traits, interests, or general intelligence are not usually appropriate to a content validation strategy. There is no rigid line however, dividing contents and constructs, particularly in the case of written tests. Although abilities are technically regarded as constructs rather than as job contents (which are behaviorally defined), it may be professionally acceptable and appropriate to validate ability measurement procedures using content validation. The crucial issue in distinguishing between the appropriateness of content versus construct validation hinges on the degree of relationship between the ability measured and job behaviors. Insofar as abilities are operationally defined, they are not hypothetical constructs and may more reasonably be considered part of the job content domain. Another way of evaluating the relationship between abilities and job behaviors involves the use of the ability in measurement. The judgment made in assessing the content validity of a test is not in terms of the degree to which applicants possess the abilities as indicated by a correlational relationship between test performance and possession of the construct ability, but is in terms of the relationship of behaviors needed in test performance to behaviors needed for successful job performance, so that test performance itself constitutes a sample of job performance. While it is inappropriate to content validate construct measures, the content validity model may provide a helpful guide to the development and establishment of the job relatedness of such instruments.

Although it is often the case that KSAOs are identified in the job analysis process, this topic is treated separately here because:

1) Test developers may have access to adequate job analysis information from which KSAOs can be derived at the start of the test development process.

2) Additional information about the KSAOs, relating to their importance, level of difficulty, and ability to differentiate is needed to properly apply the content validity model.

When a jurisdiction is beginning a test development effort without prior information about the job, it is recommended that the steps for deriving KSAOs be incorporated into the job analysis phase. Where adequate job analysis information is already available, however, test development can begin with the identification of KSAOs for measurement and the determination of their importance, level of difficulty and ability to differentiate as discussed below.

KSAO coverage. Since claims of content validity rest on a showing that the total selection program covers a substantial or critical portion of the job content universe or that tests which are components of a multipart selection program adequately sample selected job content domains, efforts must be made to generate a fairly exhaustive, yet manageable list of necessary KSAOs. Where these are derived from information about job duties, the linkage between the duties and each KSAO should be documented. The process will insure that no superfluous KSAOs are listed, and that the various aspects of the job performance are adequately covered, that is, that the elements (individual knowledges, skills, abilities, or other worker characteristics) of each domain are completely specified. Proper application of the content validity model also requires that testing exclude coverage of KSAOs which "an employee will be expected to learn on the job" (Federal Executive Agency Guidelines on Employee Selection Procedures, 1976, Sec. 12c(1)).

Determining importance of KSAOs. The content validity model allows for the restriction of KSAO coverage to those which are critical or important to job performance. Degree of criticality can be judged for each KSAO absolutely or by comparing them on this dimension in some kind of ranking procedure. Criticality can be based on the frequency, importance, or necessity (i.e., a prerequisite to successful job performance or training) of the KSAO (American Psychological Association, 1974, p. 29). Some job analysis methods incorporate this rating of importance into the technique and report only critical or important job duties. Important job duties are determined by such criteria as frequency, proportion of time spent, level of difficulty, consequence of error, or training needed to perform.

Determining level of KSAOs. Measurement of the KSAOs should not be at a higher level of competence than is necessary for successful job performance. Therefore, unless the panel of experts who will be writing and assembling the test are completely familiar with all aspects of the job, the KSAO statements should include an indication of the appropriate level of the criterion standard at which the KSAO should be assessed. For a relatively homogeneous job content domain it may be possible to specify the appropriate level of measurement for the entire domain rather than at the level of the individual KSAOs. In either case, unless these levels are directly indicated in job behaviors, the manner of their derivation must be documented.

Determining KSAOs which differentiate. If any instrument used in the selection or promotion program is to be used to rank candidates, it should be shown that the KSAO measured differentiates between lower and higher levels of job performance. That is, possession of more of the KSAOs in the content domain sampled results in better job performance. Therefore, the ability of each important KSAO to differentiate among workers must be determined.

These determinations of KSAO importance, level, and ability to differentiate are crucial to the development and validation of the measurement instruments, and must be fully documented. It is usually most efficient in terms of staff time and documentation requirements for at least the KSAO generation, importance rating, and level and differentiation determinations to

be made by the same individuals or panel of job experts.

Development of a Selection Plan

This stage of designing a selection or promotion program is not necessarily performed by the job experts involved in the job analysis phase, although their prticipation may be helpful or most efficient. It is necessary that someone knowledgeable about selection techniques be available to aid in the development of the selection plan; this person could be an examiner with relevant experience. However, if a trained psychologist is available, he or she should participate in this part of the test development process, if at all possible (American Psychological Association, Division of Industrial-Organizational Psychology, 1975, p. 10).

The development of the selection plan is important in the content validity model because it links the test to the job and documents that the test was developed specifically to measure important aspects of job behavior. Where the entire selection program is to be supported solely on the basis of content validity, the selection plan documents the scope of coverage, the appropriateness of the relative emphasis of KSAOs, and the criticality of the job content domains measured. In multipart assessment programs, however, it may be necessary to employ different validation strategies for the components of the program. In any case, the documentation of the adequacy of the coverage of important aspects of the job in the selection plan is essential to a claim of validity, regardless of the strategy employed because it establishes the job relatedness or the procedures. Figure 2 details the operations involved in developing the selection plan.

Determination of Appropriate Measurement Instruments

For each important KSAO (or each KSAO necessary for the performance of an important job duty) an appropriate means of measurement must be determined. Appropriate measurement instruments may include written tests, performance tests, training and probationary periods, physical requirements, evaluations of education and experience, interviews, application forms, scored references, or background investigations. The determination of the "appropriate" measure may be made in terms of the congruence between job behaviors and test behaviors, the expected psychometric properties of the instrument in terms of the reliability and validity of similar instruments in similar measurement situations, and the efficiency, practicality, feasibility, and relative lack of adverse impact of using the instrument in relation to other available measurement techniques. Such conditions as the resources of the employer (staff, testing time, and physical and financial resources), the characteristics of the labor market, and probable validation strategies should also be considered.

The determination of the appropriate appraisal method for each KSAO requires extensive familiarity with the wide range of instruments available. It is the responsibility of each user to study thoroughly the possibilities and make decisions, whenever possible, on the basis of empirical data which relate the proposed measurement instrument to the KSAO and indicate the technical adequacy of each instrument in the proposed measurement situation. This information, available in such sources as the personnel measurement and psychometric literature and in technical publications and manuals prepared by commercial testing firms, requires careful analysis and interpretation. This is the stage at which the services of a qualified psychologist is most necessary both as a resource person in the development of the selection program and as a technical expert in documenting the soundness of the selection plan. The basis for choosing a particular measurement instrument for a KSAO should be fully documented.

It is recommended that the selection plan be developed from a list of the KSAOs previously determined. (For documentation purposes it is also desirable to include the duties and other observable aspects of job performance associated with each KSAO, although these are generally irrelevant to the development of the selection plan.) This list should include each KSAO, an indication of whether the KSAO differentiates (is used for ranking) or is a screen-out (only indication of minimum performance ability is necessary), and the level of performance. Each KSAO should be listed only once. (The Selection

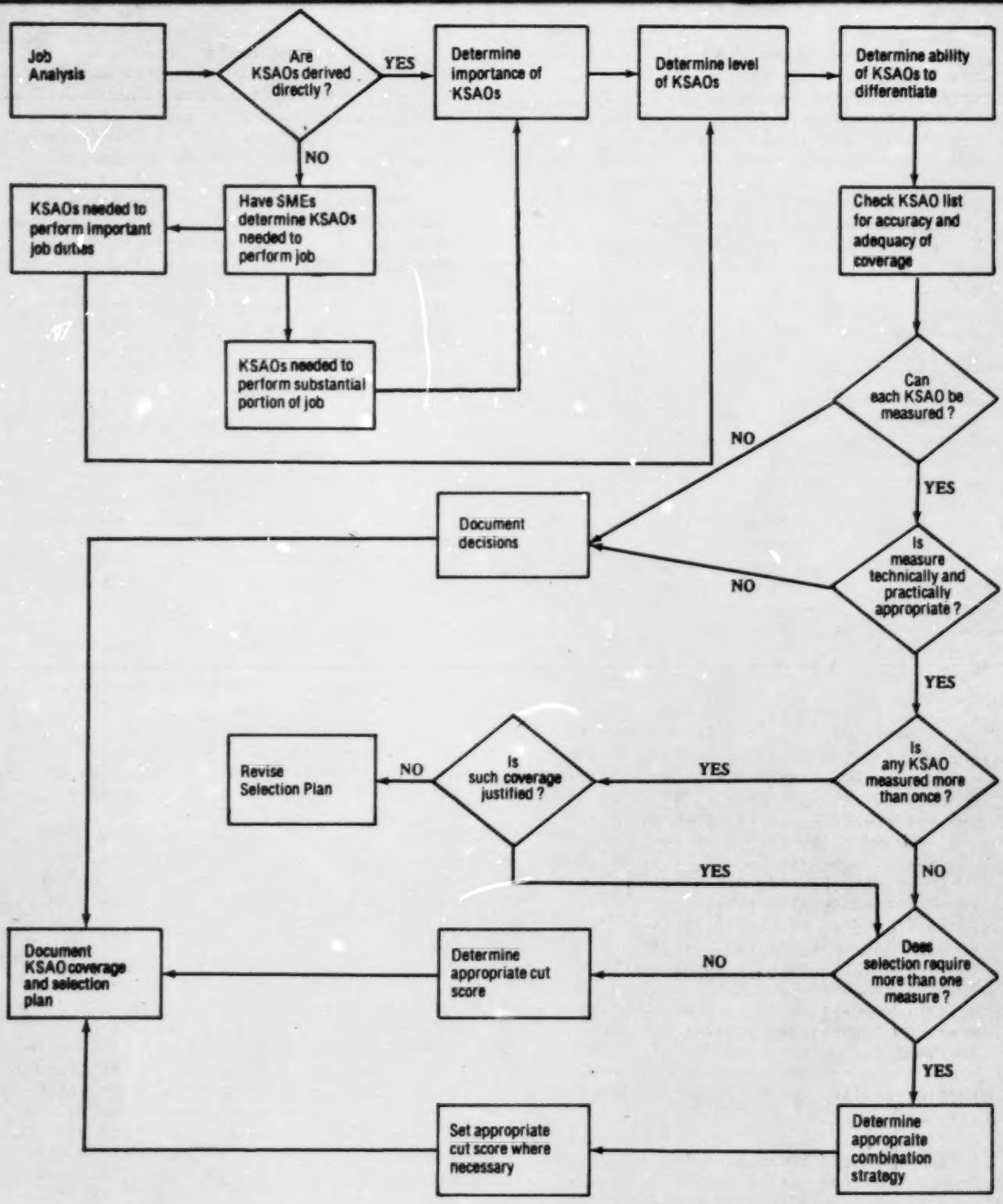


Figure 2. Sequence of Operations Involved in Developing a Selection Plan.

Plan format in Appendix A would be suitable for this purpose.) Next to each KSAO should be listed the measurement instrument decided upon. If more than one instrument is indicated for a KSAO, the panel must decide which one to use or document the advantage or necessity of double coverage and how the results will be combined in assessing the KSAO. If a KSAO is measured by more than one instrument, it must be scored in such a way that it is not given more weight in the total selection program than intended. Care must be taken to avoid developing a selection program or written test in which KSAOs which are relatively easy to measure are inadvertently given disproportionate coverage. If no instrument is indicated for a KSAO, the reasons for this lack of coverage must also be given. Although it is not necessary to measure all aspects of job performance, failure to consider important KSAOs in selection should be explained. Acceptable reasons for such omissions include lack of an appropriate instrument (including the impossibility of developing one), desire to measure only KSAOs which differentiate between levels of performance, or the fact that the KSAO can normally be learned in a brief orientation period.

It is important to remember that use of some selection devices often dictates the appropriate validation strategy. For example, a general mathematics reasoning test composed of word problems involving the calculation of distance, size and weight can be used as a measure of the KSAO "bookkeeping ability" only if it is shown to be related to that ability in a construct or criterion-related validation study, since it cannot otherwise be inferred that such general mathematical ability as measured by the test is related to the bookkeeping ability needed on the job. A test composed of samples of bookkeeping tasks such as those actually performed on the job, however, could be appropriately supported as a measure of "bookkeeping ability" by a content validation strategy.

Combining Measurement Results

After an appropriate measurement procedure has been decided upon for each KSAO to be assessed, it must be determined how the results of each procedure are to be used in arriving at a final rating. This determination involves both rational and statistical considerations.

For those instruments used as screen-outs, no combination strategy is necessary since no relative score is usually assigned, only a pass-fail rating. The rationale for the order of their presentation, however, must be documented.

When measures used for ranking are combined in a selection program, the proper weighting of each component is an important and complex concern. A discussion of the statistical issues involved in weighting components will not be attempted here, except to point out that the statistical properties of scores on measurement instruments affects the weight each score carries in the composite rating. The weighting effect of the number of items on each test and their relative difficulties can be eliminated through the use of standard scores which put the scores of each instrument on the same scale. (Standard scores are discussed in most elementary psychological statistics texts.) There are also procedures for assigning optimal weights to components which take into account such statistical properties as their intercorrelations; these procedures (regression techniques) cannot usually be applied in a content validity situation however, since they involve correlation with a criterion measure.

At issue here is the determination of the rationale for the assigned (nominal), rather than statistical (effective) weighting scheme. The panel must decide if the ranking KSAOs are each necessary to some optimal degree for job performance, or whether a low level in one can be compensated for by a greater degree or another. In other words, a decision is made regarding the use of the multiple hurdle or compensatory approaches or some combination of the two. The multiple hurdle approach requires that applicants achieve some minimum score in each measurement procedure in order to be eligible to compete in the next stage of the selection program. This approach implies equal weighting of all or of at least the essential parts of the examination, with the emphasis on the setting of appropriate cut scores for each. The rationale for these cut scores must be documented. Likewise, if the compensatory approach is recommended in which applicants receive one rating based on some combination of their scores on all aspects of the selection program, the rationale for the scoring and weighting of each instrument must be documented. Weights may be

determined by the number of important duties associated with the KSAO measured, by a special rating by job experts for this purpose, or by any other suitable and documented procedure. In many cases the weights may already be built into the selection program because of the way in which the job domain has been sampled, that is, aspects of the job are covered in proportion to their relative importance.

The format provided in Appendix A, or a similar one, should be used to provide documentation of the following:

- o the coverage of each KSAO or the rationale for lack of coverage
- o that no KSAO is covered disproportionately
- o the link between the KSAOs and the measurement instruments
- o the use of each measurement instrument
- o the rationale for the combination approach chosen (multiple hurdle, compensatory or some combination of the two) for establishing the qualifying or cutting score
- o the determination of appropriate weights (if necessary)

Preparation for Test Development

Once it has been decided to use a written test in the selection program, a test development panel should be convened.[1] This panel will be responsible for defining the test objectives and job content domain(s) to be sampled, for the development of a detailed test content outline, and for the preparation and/or selection of test items to meet the specifications of that outline. Where a content validation strategy is used to support the use of the test, such claims must rest on the manner in which the test was developed. In such cases the credibility of the entire test development process can depend on the quality and representativeness of the panel members.

Size of Panel

There should generally be three to six panel members. The amount of item writing anticipated and the time allowed for test development are factors to be considered. The minimum of three provides for some diversity of opinion. More than six members may be unwieldy and is usually unnecessary. If a great deal of item writing is anticipated, a list of additional subject matter experts should be compiled. These experts can be called

---

[1]When test items are provided by the U. S. Civil Service Commission or from other sources such as regional item banks, it is the responsibility of the using official to convene this panel. Although it may under some circumstances be possible for one person who is extremely knowledgeable about a job to develop a test using the content validity model, the recommended approach involves the participation of a panel of subject matter experts.

The USCSC provides test items and other test materials to State and local governments for some Federal grant-in-aid programs and for other occupational areas where it has developed test materials. Those test items developed and/or reviewed since 1973 will have documentation easily available as to the content domain covered and the qualifications of the subject matter specialists who wrote or reviewed the items. For older materials, such documentation can be developed only in case of essential need because of the time required to search extensive files for this information. The Commission's development of these items did not include a determination of their validity and job relatedness in each State and local jurisdiction. Therefore, the responsibility for determining the validity and job relatedness of these test materials must be determined by the user. The Commission makes no claim for their validity.

upon to submit ideas for test items for the panel to review and polish, if necessary. This procedure is recommended to speed the test development process and to provide broad coverage of the content domain which may be difficult to obtain when only a few people provide all the items.

Qualifications of Panel Members

Panel members should be supervisors, superior incumbents from the job for which the test is being developed, or other experts on the content of the job. If the job classification covers more than one division of an agency or a number of geographical locations, panel members should represent these differences. Attention should also be given to other considerations in the composition of the panel in relation to the applicant population such as race, sex, or ethnic background. All members of the panel, whether supervisors or incumbents, should have first-hand knowledge of, and preferably experience in the job for which the test is being developed, at the level for which selection is to be made. If tests are to be developed for more than one level of a job, panel members must be representative of these various levels. (Where the same panel is preparing tests for more than one job level, members must not include those who would be eligible for promotion to the higher level using the test being prepared.)

In addition to the primary requirement of thorough familiarity with the job for which the test is being developed, the panel members should also be selected on the basis of their interest, motivation, ability to work in a group situation, and potential for test development work. In many cases it is difficult to judge the qualifications of a subject matter expert in these other important areas, but whenever possible this should be done. It is especially necessary in any case, to avoid selecting as panel members those persons who are most expendable. Since a claim of content validity rests on the judgment of the panel members and therefore on their qualifications to make such judgments, it is essential that every effort be made to insure that they are in fact qualified for assignment to the panel.

Documentation File

A test documentation file should be started as soon as the decision to develop a test is made. All decisions and steps in the test development process must be supported by records of how and why choices were made. The documentation file will include all written materials relating to the project, such as the record of appointment of panel members and the completed test content plan. Records of the steps leading to the decision to use a written test should also be included in the documentation file. Such records include the documentation of the job analysis and of the selection plan as detailed in the pertinent sections of this guide. Appendix B lists in summary form the materials which should be included in the documentation file.

A resume of each panel member's relevant experience and education should be prepared for the file. These resumes should include such information as dates of relevant experience, dates of promotion, and description of job duties. The panel members' current positions should be described in detail. If a measurement specialist is involved in the test development process, a resume of his or her qualifications and experience should also be included in the documentation file.

Training

When test development requires item writing it is often necessary to provide a special training session for this task. Although panel members are assumed to be thoroughly familiar with the content domain to be sampled, guidance concerning the editorial and technical requirements for good item writing is usually necessary. A measurement specialist familiar with these requirements can be helpful in providing such training and in reviewing and revising the completed items.

Security

Care must be taken to maintain the security of all materials relating to the test development. The documentation file, job analysis data, selection plan, test outline, and test questions should be kept in a locked file cabinet. Panel

members should be informed of security requirements and cautioned not to speak of any of the work in progress to others not on the panel. Security forms (see sample form, Appendix C) should be signed by all persons working on the test and these forms filed with the documentation materials.

Security is enhanced if the panel is provided a working space large enough to accommodate all members in one room. The room itself should be so situated that there is no chance that conversations concerning individual test items will be overheard. No test materials should be left unattended. At the end of each working session, all test materials and related documents should be locked in a secure file cabinet.

Development of the Test Content Outline

The test content outline (sometimes referred to as the test plan or test specifications) links the test to the job content domain it samples and is therefore an essential piece of evidence in the demonstration of content validity. The explicit definition of the job content domain to be sampled is a necessary part of the documentation of this link.

Definition of the Job Content Domain

In the selection plan, the definitions of the job content domains are implied in the listing of the KSAOs associated with each measurement instrument. In the case of written tests, this listing will usually be in terms of knowledges or skills. Any conditions applying to these individual knowledges or skills (such as, the ability to differentiate, or importance) will be part of the definition of the job content domain. In explicitly defining the domain from the information in the selection plan, it may be helpful to organize the knowledges or skills into broader categories for ease in test development.

The explicit definition of the job content domain for written test development will therefore consist of an organized list of the knowledge or skill elements[1] comprising the domain and a descriptive statement delineating the boundaries of the domain, such as, "critical job knowledges," "skills frequently needed on the job," or "knowledges differentiating better from poorer workers." The definition of the job content domain must be complete and explicit enough to support judgments that:

1) the behaviors elicited by the measurement instruments correspond to behaviors needed on the job; and,

2) the measurement instrument adequately samples the elements of the relevant job content domain.

Definition of Test Objectives

The definition of test objectives should make explicit what is implicit in the selection plan. Definition of the test objectives takes into account exactly what is to be measured, the level at which it is to be measured, the purpose of measurement, the method of measurement, and how test results are to be interpreted. Test objectives set the standard against which the success of the testing program can be judged. A suitable statement of a test's objectives might read, for example, "to measure by means of a written test knowledge of elementary statistics needed in test development work in order to rank competitors."

Definition of a test's objectives is important in focusing the test development efforts. It determines what kind of a test (e.g., job knowledge test, performance test, aptitude test) is to be used and helps in

---

[1]This discussion of job content domains is in terms of those relevant in the development of written employment tests. For work samples and other types of performance tests, however, the job content domain may be defined in terms of such things as "work behaviors, work products, work activities, [or] job duties" (Federal Executive Agency Guidelines on Employee Selection Procedures, 1976, Sec. 12c(1)).

deciding the format of test questions. As mentioned in the discussion of validation strategies, the objectives of the test determine the appropriate validation method insofar as the test objectives define the inference to be drawn from test scores (e.g., prediction of future performance or measurement of present performance). Furthermore, the objectives explicitly state the way in which the test is to be used (e.g., for ranking or as a screen-out) as was determined by the analysis of the KSAOs.

Practical considerations may also be taken into account when formulating test objectives. For example, if a large number of applicants is anticipated, the test objectives might include a statement that it is to be used as a screen-out (i.e., used on a pass-fail basis) for further consideration in a ranking process. Often, there are regulations or established promotion plans which dictate aspects of the test objectives.

Planning the Format, Timing and Length of the Test

Decisions regarding the appropriate format of test questions, the length, and timing of the test are part of the development of the test content outline.

Format of test questions. The format of the test questions may be established in regulations or may be established by custom. If the format is not prescribed by regulation, the panel should make the choice in terms of ease and accuracy of scoring and professional appropriateness of measurement.

If the items are carefully constructed, the multiple-choice format is probably the most appropriate choice for an objective written test. In deciding upon the number of alternatives in the items, both the chance level of answering the item correctly and the length of the total test should be taken into account. Although items with only three or four alternatives are more likely to be answered correctly by chance by an applicant who does not really know the answer, it may be difficult to write five good alternatives. If the test contains a sufficient number of items, the effects of such guessing will not be too great; for shorter tests, however, the additional effort needed to write five alternatives may be necessary to insure that test scores adequately reflect applicants' true abilities.

The true-false question format, while easily scored, is more susceptible to the effects of guessing and therefore does not differentiate among candidates as well as the multiple-choice format. Open-ended questions, essays, or fill-in-the-blank questions can provide good measures of knowledge, but are not easily and reliably scored. The answers are subject to interpretation by the scorer, and some answers may be "more right" than others.

Timing of test. The timing of the test is dependent upon its length. There should be enough time allowed for all except an extremely slow candidate to consider and answer each question. (Nunnally (1967) suggests that 90% of the candidates should have enough time to complete the test.) Some questions are more complex than others and require more time. Without any pretesting, it is not possible to know exactly how long the candidates will take to complete the test. However, allowing an average of one minute for multiple-choice items not requiring any computation and 2½ to 3 minutes for questions requiring some computation or reading of complex materials is probably more than satisfactory. The availability of pretest data or information on the time needed to complete comparable items on other tests allows for more precision in setting time limits and is generally preferred to the use of such rough estimates.

In tests over 1½ - 2 hours, attention should be given to the issue of fatigue and boredom in applicants. If a test requires 3½ or more hours to complete, it is advisable to provide a break.

Length of test. The principle that governs the number of items in a test is that there must be enough to show the degree to which each applicant possesses the elements of the content domain in comparison with each other applicant. The adequacy of this showing depends to some degree on the complexity of the job content domain being measured. There should be enough questions

to sample adequately all the knowledges or skills defined as belonging to the job content domain the test is sampling.

The length of the test can be set at a fixed number of items in order to simplify scoring or to be consistent with regulation requirements. It is not necessary to do this, however. The length of the test can be established as the test content outline is developed, making the number of test items equal to the number of knowledges or skills present in the job content domain or allocating a certain number of items to the measurement of each KSAO. If a fixed number of items is desired, experience indicates that a range of 80 to 120 items is generally appropriate. This range usually provides enough items to reliably sample a job content domain composed of a set of fairly complex knowledges or skills without wasting additional test development and applicant time. In determining the appropriate number of test questions, administrative considerations must also be taken into account. The cost of constructing, administering and scoring a test tends to increase as the length increases, and these factors should be weighed against the psychometric considerations regarding test length.

In summary, a test should include as many items as are necessary for coverage of the job content domain but few enough so that the test is practical to administer and within the physical capacities of the applicants.

Planning the Content Area Coverage of the Test

The test to be developed must adequately sample the job content domain. The method of this sampling must be specified. Usually claims of adequate sampling are based on the relative number of test questions measuring each individual knowledge or skill or groups of knowledges or skills in the job content domain. Ideally, the test should measure each element of the job content domain in proportion to its importance to that domain.

Determination of the appropriate proportion can be made in a number of ways, depending on the type of job analysis information available and the test objectives. For example, if a particular job knowledge is to be measured in a test covering the job content domain of all important job knowledges, its relative importance in the total test might be determined by one of the following methods:

1) Count up all the knowledges in the domain (N) and use the value 1/N as an index of its relative importance. (This method assumes all job knowledges are equally important to job performance.)

2) Find out how many important job duties are associated with each knowledge and use this value as an index of its relative importance.

3) Obtain a consensus by panel members of the appropriate weighting for each element of the job content domain.

4) Randomly sample from the domain. (This is not generally recommended for content valid tests (American Psychological Association, Division of Industrial-Organizational Psychology, 1975, p. 10).)

It is often helpful, if KSAOs have been grouped into broader categories, to determine the relative proportion of items needed for these broader groups. The individual KSAOs in each group can then be sampled according to the number of items assigned to the category. This method is also helpful in avoiding overemphasis of an area which may encompass many KSAOs but is not overly important in the job content domain. For example, in a test for test technician the broad job content area "Knowledge of statistics necessary for test construction" may include five particular knowledges, but be more important than the broad area "Ability to perform simple statistical calculations" which includes eight particular knowledges and abilities. Although grouping of KSAOs is often desirable, such a step may be necessary when the number of KSAOs exceeds the number of test items.

In the content validity model it is necessary to document the method of sampling used and the adequacy of the test as a sample of the job content domain. The test content outline serves as this documentation.

The Test Content Outline

The test content outline is basically a summary of the procedures detailed above; that is, a list of the knowledges or skills in the job content domain with an indication of the number of test items devoted to each. The list of knowledges or skills may be grouped into broader categories. The outline should be detailed enough to guide the selection of items and test assembly and to document the adequacy of the sampling procedure. A sample test plan is given in Appendix D. The test plan should be retained as part of the documentation file.

The test content outline should be carefully reviewed by the panel. The members should agree that the outline adequately covers the job content domain. If the outline does not reflect the domain, the panel should carefully examine the definition of the job content domain and sampling procedures to ascertain any problems. This does not mean that the outline should be changed out-of-hand. It was developed from the pooled judgments of the panel members, using a carefully designed procedure. Any changes made in the test content outline should be made with care, and reasons for the change thoroughly documented. Once the test content outline is approved by the panel, it should be changed only if it is discovered that the job content domain was described improperly. In that case, the entire process of developing the outline should be carefully reviewed.

Test Construction

The actual test construction involves assembling the proper number of appropriate items in each content area to meet the specifications of the test content outline. Items may be taken from existing files or written specifically for the test. All items should be carefully reviewed before the test is put into final form.

Item Writing

New items should be written using techniques for good item writing to measure the KSAOs indicated on the test content outline. They should follow the format agreed upon for the test items and be at the suitable level of difficulty. Discussions of good item writing techniques are given in test development and measurement texts such as Ebel (1972) and Thorndike (1971). A guide to item writing is given in a separate short report Item Writing Guide, Sample Items, and Examples of Item Modification by Lynnette B. Plumlee (1976), and is available from the Personnel Research and Development Center upon request.

Criteria for Item Review

All items, whether from the jurisdiction's file, the USCSC, other outside sources, or newly written, must be reviewed by the panel members before they are put in the test. This review should consider the technical and editorial adequacy and accuracy of the items and their relevance to the content area being measured. Reviewers should also make sure the items do not use language that is more complex and technioal than that used on the job. Members should review the items individually, and then meet as a group to resolve any problems.

Items should first be reviewed to determine their relevance to the content areas being measured. Items not measuring any of the content areas from the test content outline should be dropped from further consideration. New items should also be subject to this review to determine that the final items measure the areas they are supposed to measure. It is helpful to code the items according to the knowledge or skill measured for ease in test assembly and for filing purposes. This review should include a determination that each item measures important, non-trivial aspects of the knowledge or skill.

Once it has been determined that the item is relevant to the test content, the answer key must be carefully reviewed by the subject matter experts on the panel. For most types of objective test questions there should be one and only one correct answer. There are some objective item types to which this rule does not apply. These include questions which require competitors to rank the alternatives, or questions in which each alternative is given a relative weight in assigning test scores. In all cases, however, the source of the answer key should be documented.

The item should present only one problem in a clear, specific and accurate manner. Since the test is designed to distinguish better applicants from less able ones, reviewers should try to deter-

mine if the correct answer is likely to be chosen by more of the better applicants. Conversely, wrong answers (distracters) should be plausible and appeal to the less able applicants. The item should be written at the appropriate level of difficulty. The level of language used in the item should be similar to that used on the job.

Reviewers should be sensitive to any possible bias in the item. Is the item testing an obscure or irrelevant aspect of the knowledge or skill? Would a person be unable to answer correctly because of ambiguous or overly difficult wording of the item? An affirmative answer to any of these questions indicates that the item should be revised or discarded.

Once the content and technical adequacy of the test items have been determined, they should be given an editorial review. Items should be checked to make sure they are in the proper format, are correct in terms of English usage, punctuation and spelling, and are free of colloquial expressions. Care should also be taken to insure that the items do not use language with sexist bias (e.g., unnecessary use of sex-specific pronouns) or present situations in terms of stereotyped sex roles. Items should be stated clearly and concisely, free of unnecessarily technical or complex language and unintended hidden clues as to the correct answer.

Each panel member should be instructed to discard any item believed to be inaccurate, to have more than one correct answer, or to be unfair or trivial. Specific corrections, additions or other improvements should be suggested in writing.

After the individual panel members have completed their reviews, the panel meets as a group to resolve their problems with the test items to determine the pool of items in final form. The items in this pool should then be given a final review by panel members.

It is recommended that each item in the final pool be typed on a card for ease in test assembly and to serve as a record of item documentation. Panel members should record their initials directly on this card when they are completely satisfied with the item. The initials signify that the panel member believes the item to be significant, relevant, clearly stated, fair, able to differentiate between the best and poorest applicants, and correct to the best of his or her knowledge. The meaning of the individual initials on each item card should be clearly explained in the test documentation.

No items are to be considered for inclusion in the test unless the item card is initialed by each of the panel members. When all of the items have been individually reviewed, test assembly according to the test content outline is begun.

Test Assembly

The choice of specific items to be assembled into the final test is made based on the content of each item, the comparative quality of items written in the same content area, and the appropriateness of the estimated or actual item difficulty levels. As a general rule, it is desirable to have from 1/3 to 1/2 more acceptable items to select from than called for in the test content outline. If the item file is found not to contain enough items to sample a content category as specified in the test content outline, new items should be constructed and put through the same reviewing and approval process as the items already in the item file.

Once the final set of items has been chosen by the panel, the item cards are stacked in the same order as the questions are to appear in the test. It is generally preferable to group items referring to similar knowledges or skills together in the test.

The ordered item cards, which represent the final complete test, are then reviewed by the panel members. The purpose of this review is to check for inconsistencies of reference in different items and possible answer give-aways between items because of content.

The number of correct answers which are keyed A, B, C, and D (and E if used) should be tallied for the total test. Each letter choice should be the correct answer about the same number of times (i.e., within 10%), and the correct answer letter choices should appear in an essentially random pattern. If necessary

to meet either of these conditions. the alternatives of some of the items should be rearranged if possible. Other solutions (such as writing new alternatives or replacing the item) may be necessary when the order of the alternatives cannot be altered; arithmetic reasoning items, for example, typically present the alternatives in ascending numerical order. An overuse of one correct letter choice or of an answer pattern can provide clues to applicants in selecting their answers to questions they do not actually know.

The test copy is then typed from the selected and ordered item cards or otherwise prepared for duplication. The final copy should be proofed twice by the panel members. An official answer key is prepared, and it also must be proofed twice. Each panel member (or selected members) then reviews the test again, this time as a complete unit, and prepares a key by answering the items. These keys are used to verify the official key.

When all corrections or changes have been made on the test copy and answer key, an official file copy is made. This copy is initialed by each panel member, attesting that he or she has reviewed the final test, and judges it to be fair and job related, adequately sampling the important and differentiating KSAOs belonging in the job content domain.

The item cards and initialed file copy of the final test serve as documentation. Additionally, a summary sheet should be drawn up showing each item number, its source and the content area it measures. This sheet should also be maintained in the documentation file.

Pretesting

It is often helpful, particularly when using a new or unusual type of test item or initiating a new selection program, to try out the items on a sample of applicants. It is sometimes better not to use the actual test questions, in case they should be compromised during the tryout phase, but instead to use a similar form of the test or a shortened test. The value of such a procedure is based on the degree to which the items pretested are representative of the actual test items. Since the development of comparable materials for pretesting is technically complex, it is preferable, in most cases, to obtain information on the actual test items. Such pretesting should be conducted only under the strictest security precautions, however.

Pretesting provides valuable information on the difficulty of the items and the relative difficulty of the content areas. The process also indicates where revisions are necessary by pointing to distracters which are not attracting applicants and to items which are ambiguous or not differentiating. Pretesting also allows for more accurate estimates of how long it takes applicants to complete the test items.

The availability of pretest data can greatly facilitate the test assembly process by providing information on the difficulty and value of each item in discriminating between the better and poorer applicants. On the basis of such data it is possible to estimate the psychometric characteristics of the final assembled test and thus document the technical soundness of the test assembly process. These procedures can also insure the comparable coverage (in terms of difficulty and score range) of various content areas in a heterogeneous test. For these reasons, pretesting is highly recommended for inclusion as a routine part of the test development process. It should be noted that many commonly recommended test assembly procedures using item data are appropriate only for homogeneous tests; insofar as content valid tests are designed to be representative samples of heterogeneous aspects of the job, item data reflecting internal consistency (such as item-test correlations) may be inappropriate and should be used with caution in test assembly.

If several different forms of the same test are to be used interchangeably with a common register (list of eligibles), then every question must be tried out in advance and the various test forms must be constructed to be parallel both in terms of content coverage and item and test statistical parameters. The details of this type of test construction are complex and not covered in this guide.

Test Administration and Scoring

A professionally developed test is of use to an employer only if it is administered fairly, in a standardized manner, and scored correctly. The following points should be taken into consideration to insure that the testing program is administered in a consistent manner, enhancing its reliability.

Test Administration

The test should be administered in a room in a relatively quiet location, with enough space provided between the applicants' seats (so that they are comfortable and so that they cannot copy from one another), with adequate lighting and ventilation, and with a good writing surface for each applicant. The time and date of the test should be set enough in advance so that all applicants can appear. It is often helpful to prepare pamphlets of sample questions to insure that all applicants are equally familiar with the types of items on the test. Sample questions may be essential in demonstrating the equal treatment of all applicants.

Specific directions for administering the test should be prepared by the panel or personnel officer. The person(s) who give the test should follow these directions exactly as they are shown. If the test is given in more than one room or if there is more than one administration, there should be no difference in the directions given.

Scoring

Generally tests developed by the content validity model are power tests in which applicants are given ample time to complete all the items. In such tests a correction for guessing is usually unnecessary since it does not change the rank ordering of the scores. Therefore, a "rights only" scoring procedure is usually used.

Scoring can be done either by hand or machine, depending on the volume of tests and the equipment available. It should be done as quickly as possible after the test date to facilitate the timely development of the list of eligible applicants. All hand scoring should be done twice to avoid errors; machine scoring should also be checked for accuracy (for example, by periodically rescoring a sample of answer sheets).

If there is a large enough group of applicants, it may be useful to perform an item analysis on the data. If the applicant group is very large, such analysis may be done on an appropriate sample of answer sheets. Inspection of such item analysis data before final scoring and official notification of scores is recommended. The data will usually point out an item which was incorrectly keyed, because an unusual proportion of high-scoring applicants will be shown as getting the item wrong. The item analysis review may also highlight items which were constructed poorly, and were unclear or ambiguous. It is up to the user to decide whether such an item should be dropped from the total score of all applicants, of if some scoring adjustment (such as scoring two alternatives as correct) should be made. Any such scoring exceptions should be made before applicants are notified of their test scores.

Setting Cut Scores

Unless regulations require the setting of a minimum passing point (cut score) or a multiple hurdle approach to combining measures, it is not necessary to set a cut score provided there are enough qualified applicants with high test scores to fill current openings. A cut score is unnecessary in such cases since the test is used only to rank applicants and the test score is combined in a compensatory approach with other factors to determine the final rating.

If a cut score is to be set, it is essential legally and professionally to document the procedures and reasons for choosing the particular passing score. There are many factors which may be taken into account when setting cut scores.

Most guidelines on testing and selection and some legal cases have affirmed that cut scores should be "reasonable and consistent with normal expectations of acceptable proficiency within the work force" (Federal Executive Agency Guidelines on Employee Selection Procedures, 1976, Sec. 5f). This implies that the passing point should be set at a level which corresponds to some minimum level of performance. This point may be determined

empirically (for example, by testing present employees and establishing employee norms) or rationally (for example, by determining the minimum number of KSAOs needed for successful job performance).

There are other factors in addition to minimum performance level which may also influence the setting of cut scores. One of these factors is the selection ratio; i.e., the number of hires or promotions to be made in relation to the total number of applicants. Generally, the higher the selection ratio (that is, the fewer applicants there are relative to the number of openings), the lower the test cut score can be.

When taking administrative factors into consideration employers should carefully evaluate the degree of any adverse impact. (Adverse impact refers to the situation in which a given cut score tends to reject disproportionately more members of one group than of some other group or groups.) The compilation of applicant norms for the test are often useful in assessing adverse impact and setting reasonable cut scores.

The test user may also find it helpful to take such factors as the test's reliability and standard error of measurement into consideration in determining the cut score to minimize the effects of any chance differences among applicants. The total number of measurement instruments used in assigning ratings may be another consideration. If a number of measurement instruments are used, for example, the cut score on each should be set so that minimally qualified applicants will be eligible for the next part of the selection process and not eliminated from further consideration on the basis of an unnecessarily high cut score. Also, test users should consider the economic and human risks of hiring an unqualified applicant relative to the risk of rejecting a qualified applicant.

Regardless of how the cut score is determined, it should be above the score that applicants might be expected to achieve by chance. Generally, for carefully developed multiple-choice tests the chance score is assumed to be the number of items on the test divided by the number of response options for each item.

The essential fact to remember in setting cut scores is that there must be some "rationale, justification, or explanation of the cutting score adopted.... The intent is to recommend that test users avoid the practice of designating purely arbitrary cutting scores they can neither explain nor defend" (American Psychological Association, 1974, p. 66-67). In addition to the professional considerations, unsupported, arbitrary cut scores may be subject to legal attack.

Technical Adequacy of Test

If properly followed, the procedures discussed in previous sections will insure the development of a test for which content validity reasonably can be inferred. This section describes some additional information which is highly desirable to obtain. For example, there are statistics which are necessary to interpret properly the meaning and value of particular test scores. These include at a minimum, a frequency distribution, a measure of central tendency, and a measure of variability of the test scores. (These statistics are used in setting cut scores and in developing formulas for translating test scores into ratings, and should therefore already be available in most cases.) The reliability of the test scores should also be determined where feasible and appropriate to provide an estimate of the stability of the test scores. These statistics should be retained in the documentation file to support claims that the test is being properly used in the selection program. Test users may also find it useful to analyze the item responses to learn how each item contributed to the total test score and to make scoring exceptions (e.g., not score an item if the data indicate that the item is not discriminating appropriately between high and low test scorers and the documentation for the item is weak). These statistical measures are discussed briefly in the following sections. Specific formulas required for computation are in most cases not given; these can be found in basic statistics books.

More sophisticated and extensive statistical analyses are often required if tests are legally challenged. It is recommended that the applicants' answer sheets (or a record of them suitable for computer analysis, such as punched cards or magnetic tape) be filed with other documentation materials so that these and any other additional statistical calculations

and analyses can be made later, if necessary. These should be retained for at least as long as the test is in use. (If there are a large number of answer sheets and limited storage space, a random sample may be chosen for retention. In this case, a minimum of 400 answer sheets should be saved. It should be noted, however, that jurisdictions are often required to retain all answer sheets through an appeals period and usually for the life of the list.)

Frequency Distribution

The basic information needed for test analysis is found by preparing a frequency distribution table. The first column in the table lists each of the scores obtained by applicants on the test, from lowest to highest. Next to each score is listed the number of applicants receiving each of the scores. A third column contains the cumulative frequency of the scores as they add up. A fourth column gives the cumulative percentage of applicants at each score point. The frequency distribution gives a general picture of the spread of test scores, how high or low the scores tend to be, and if there are any scores which are far out of line with the rest of the scores. The information in the frequency distribution table is used to compute some of the test statistics described in the next sections.

Measures of Central Tendency

A statistic that represents the typical, or average, score of a group of scores is called a measure of central tendency. Probably the most familiar of these measures is the mean score. This is simply the average of all of the test scores. Another way to represent the typical test score is to find the point on the score scale that separates the top half of the group from the bottom half. This is called the median.

The mean and median may not correspond exactly, but usually will not differ greatly unless the set of scores is very "skewed." This means that there is a clustering of scores on one side of the mean score and a wide spread in scores on the other side. This could result if the test was either very difficult or very easy. The median is less affected than the mean by the few cases at the end of the wide spread, and therefore will give a more accurate picture of the central tendency if the distribution of scores is not even.

Measures of Variability

When describing a set of scores, it is is necessary to report how variable the scores are, i.e., how much they spread from high to low. The simplest description of spread is the range of scores in the group. This is simply the difference between the highest and the lowest score obtained on the test or a statement of the minimum and maximum scores. Since the range depends only on the two extreme cases in the total group it can be very misleading as an indication of variability, as it can be greatly affected by the addition or omission of a single extreme case.

A better measure of variability is based on the individual score distances from the mean score, and is called the standard deviation. The standard deviation indicates the degree to which test scores vary about the mean score and is based on the difference (deviation) between each test score and the mean score as follows:

1) Square each of the deviations (to remove the negative sign for scores below the mean).

2) Add the squared deviations and divide by the number of cases.

3) Find the square root of the results of step 2.

The resulting measure of variability, known as the standard deviation, increases in direct proportion as the scores spread more widely around the mean. It not only gives an indication of the variability in the group, but is useful in comparing test variability from different groups because the standard deviation has comparable meaning from one group to another. When different tests are of equal length or have results expressed in standard score units, the standard deviation has comparable meaning from one such test to another.

Measures of Reliability

The reliability of a test is basically its consistency or dependability of measurement. Reliability implies that the score a person receives on the test will remain somewhat constant over time, and that the individual will stay in about the same place in the group if another measure of the same characteristics is used or if the same measure is used at different times.

The stability of test scores is affected by variation among test forms, variation in applicants' ability between test administrations and imprecision and inadequacy in the test itself. There are three generally accepted methods of estimating the reliability of a test. These are the test-retest method, the parallel forms method, and the internal consistency method. In the test-retest method, applicants are given the same test on two different occasions and their scores correlated to determine the degree to which the scores from the different occasions are related. In the parallel forms method applicants are given two equivalent forms of the test and their scores on the two forms correlated. There are a number of ways to obtain internal consistency estimates, all of which basically involve the correlation of applicants' scores on one half of a test with their scores on another half of the same test, and an estimate of what the correlation would be if scores on the entire test had been correlated. The Kuder-Richardson formulas for estimating internal consistency result in estimates of the average correlation of all possible split-halves of the test.

Each of these methods is sensitive in varying degrees to the different factors affecting the stability of test scores, and each involves methodological difficulties. Therefore, the appropriate use of each method requires that certain assumptions be met. For example, the reliability of a content valid test is defined primarily in terms of the adequacy with which the test samples the job content domain. For this reason, measures of internal consistency are not necessarily appropriate for content valid tests since it cannot be assumed that such tests are sampling an homogeneous job content domain. It is the responsibility of the test user when reporting reliability to decide which estimate is appropriate for the particular test being analyzed.

Item Analysis

Much useful information about each test item can be found by analyzing the individual applicant responses. The basic item analysis is a tabulation of the responses that have been made to each item on the test. A tally is made of how many chose each of the possible wrong answers and how many omitted the item. It helps to have this information separated by high-scoring and low-scoring halves (or other proportion) of the group. From this type of tabulation such questions can be answered as how hard the item is, whether or not the item discriminates between the better and poorer applicants, and whether the test item was constructed well or poorly. These data are usually posted on the item cards for documentation purposes and to aid in future test assembly.

These questions can be answered in some cases by inspecting the tabulation of data. Statistical indexes of item difficulty and item discrimination can also be computed. Programs are available to prepare complete item analyses by computer, or they can be prepared by hand. Care should be taken in interpreting item data because the statistics commonly applied are appropriate for homogeneous tests whereas content valid tests are often designed to cover heterogeneous content areas.

Although inspection of the data can be very useful, it should be noted that there are many opportunities for chance to operate in preparing item analysis statistics. Unless there are many persons taking the test, the results can be misleading and should be used with some caution.

Review of Selection Program

Tests developed using the content validity model are related to job requirements and are designed to measure presently possessed knowledges, skills, abilities, or other worker characteristics. Changes in the labor market which result in a different applicant pool and/or changes in the job can affect the difficulty of the test items and the meaning of the test scores. Therefore, the testing program should be reviewed at periodic intervals to insure its continued validity and job relatedness.

This review should begin with the job analysis phase and the accuracy of the original information to describe the present job. The job duties, work conditions, and other factors, such as technological changes, should be carefully considered in this review. The selection plan should also be reviewed to insure that the measurement instruments being used are still the best available for measuring the KSAOs. The resources of the employer to develop and administer additional and/or more appropriate instruments should be taken into account.

The test itself must be carefully reviewed for currency and accuracy. The test outline, individual items, and the test as a whole should be considered. Scoring procedures should also be reviewed, as well as the difficulty level of the items. Changes in the relevant labor market may affect the item and test statistical parameters.

Records of the periodic review should include the date, names and qualifications of the reviewers, and the results of the review. These should be kept in the documentation file.

Glossary

Alternative: one of several answer options on a multiple-choice test question.

Construct: a conception that is formulated in order to account for patterns and consistencies in behavior. Constructs are used to organize complexes of observable events into categories, usually known as traits. A construct represents a hypothesis that a variety of behaviors will correlate with one another over studies of individual differences and/or be similarly affected by experimental manipulation; the construct is understood or inferred from this network of interrelationships.

Construct validity: the degree to which scores obtained on a test may be interpreted as measuring or as related to a specified construct.

Content validity: the degree to which scores on a test may be accepted as representative of performance within a specifically defined content domain of which the test is a sample. In employment situations the concept of content validity has been extended to include the degree to which a selection program samples critical job domains or domains which constitute a substantial proportion of the job.

Criterion-related validity: the statistical statement of the existence of a relationship between scores on a test (predictor) and scores on a measure of performance (criterion).

Cut score: a score at or above which applicants are eligible for selection; those below the score are eliminated from further consideration.

Difficulty index: any of a variety of indexes used to indicate the difficulty of a test question. The percent of some specified group answering an item correctly is an example of such an index.

Discrimination index: any of a variety of indexes used to indicate the extent to which a test item differentiates among examinees with respect to some standard. An example of such an index is the item-test correlation which indicates the extent to which those with high scores answer the item correctly and those with low scores answer it incorrectly.

Frequency distribution: a tabulation of scores (or other data) from high to low, or low to high, showing the number of individuals who obtain each score or fall within each score interval.

Item: a test question or problem, including, in multiple-choice tests, the set of options or alternative answers.

Item analysis: a detailed description of how an item or group of items functioned when used in a test. This description is usually given in statistical terms and should provide information about the difficulty of the item for the sample on which the analysis is based (difficulty index), the relative attractiveness of the options, and how well the item discriminated among the examinees with respect to a chosen criterion (discrimination index).

Job analysis: a systematic study of a job and of the behavior of workers in that job. For selection purposes a job analysis should provide information needed to identify the worker qualifications (KSAOs) needed for job performance.

inferences are to be made. Job content domains include work behaviors, work products, work activities, job duties, or the knowledges, skills or abilities shown to be necessary for the performance of the duties, behaviors, activities or the production of work. The validity of inferences made about job content domain depends on the adequacy with which test samples the domain, that is, on the degree to which the test content domain is the same as the job content domain. Job content domains may also be referred to as job performance domains.

Job content universe: the total job; everything, known and unknown, which the employee does and must know in order to perform the job.

Job related: a test is job related if it samples knowledges, abilities, skills, or other characteristics shown through a careful job analysis to be necessary or important for successful performance of a job.

Mean: the sum of a set of scores obtained by a number of persons divided by the number of scores; the average of a set of data.

Reliability: the consistency with which a test measures or the degree to which repeated measurement of the same individual would tend to produce the same results.

Selection plan: specification of the instruments to be used to measure applicant qualifications relevant to the various aspects of the job, including the objectives of such measurement and how the components of the selection program will be combined in assigning final ratings of eligibility. The selection plan provides a detailed description of the procedures to be used in making selection decisions and should include descriptions of the aspects of the job to be covered. The selection plan may also be referred to as the examination plan.

Selection program: any measure, combination of measures, or procedures used as a basis for any employment decision. The selection program includes all information upon which final ratings of applicants are based and often includes a variety of individual tests as components in order to provide broader coverage of job requirements. The selection program is often referred to as an examination.

Standard deviation: a measure of the variability of a distribution of scores about its mean.

Standard score: a converted score expressed in terms of standard deviation units (or multiples thereof) and related to the mean score on a predetermined scale.

Test: any procedure used to provide information about individuals. Tests include the full range of assessment instruments from traditional paper and pencil tests, performance tests, training programs or probationary periods and physical, educational and work experience requirements through informal or casual interviews and unscored application forms. Any component of a selection program is a test.

Test content domain: the performance domain which the test measures. The test content domain may be a set of behaviors sampled by test items or objectives reflected by test items. The test content domain is specified in the test content outline.

Test content outline: specification of contents of a test indicating the content areas covered, the number of questions in each area, the format of test questions and the length and timing of the total test.

Validation: the process of investigation through which the degree of validity of a test or a selection program can be estimated. The research (validation) process should not be confused with the results of such research (validity evidence) which may or may not justify the use of a test for a particular purpose.

Validity: the degree to which certain specified inferences from scores on tests may be considered justified.

References

American Psychological Association. Standards for education and psychological tests (Rev. ed.). Washington, D. C.: Author, 1974.

American Psychological Association, Division of Industrial-Organizational Psychology. Principles for the validation and use of personnel selection procedures. Dayton, Ohio: The Industrial-Organizational Psychologist, 1975.

Ebel, R. L. Essentials of educational measurement (2nd ed.). Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Equal Employment Opportunity Commission. Amended, 29 CFR 1607, Guidelines on employee selection procedures. Federal Register, 1976, 41, 51983-52051.

Federal executive agency guidelines on employee selection procedures. Federal Register, 1976, 41, 51734.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Office of Federal Contract Compliance Programs. Amended, 41 CFR 60-3, Guidelines on employee selection procedures. Federal Register, 1976, 41, 51744-51751.

Plumlee, L. B. Item writing guide, sample items, and examples of item modification. Washington, D. C.: U. S. Civil Service Commission, Personnel Research and Development Center, 1976.

Prien, E. P. & Ronan, W. M. Job analysis: A review of research findings. Personnel Psychology, 1971, 24, 371-396.

Thorndike, R. L. (Ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

U. S. Civil Service Commission. Appendices to Federal Personnel Manual Supplements 271-1, 271-2, 335-1, 990-1 (Book 3), Part 900, Subpart F, Employee selection procedure guidelines. Federal Register, 1976, 41, 51752-51759.

U. S. Department of Justice. Amended, 28 CFR I-50, Guidelines on employee selection procedures. Federal Register, 1976, 41, 51735-51743.

Wright, G. H. Public sector employment selection. Chicago: International Personnel Management Association, 1974.

Bibliography

Adkins, D. C. Test construction (2nd ed.). Columbus, Ohio: Charles E. Merrill, 1974.

American Psychological Association. Standards for educational and psychological tests (Rev. ed.). Washington, D. C.: Author, 1974.

American Psychological Association, Division of Industrial-Organizational Psychology. Principles for the validation and use of personnel selection procedures. Dayton, Ohio: The Industrial-Organizational Psychologist, 1975.

American Psychological Association. Guidelines for non-sexist language in APA journals. American Psychologist, 1977, 34.

Biddle, R. E. Guidelines oriented job analysis (3rd ed.). Sacramento, California: Author, 1976.

Byham, W. C. & Spitzer, M. E. Personnel testing: The law and its implications. Personnel, 1971, September-October, 8-19.

Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper & Row, 1970.

Cronbach, L. J. & Gleser, G. C. Psychological tests and personnel decisions. Urban, Illinois,:: University of Illinois Press, 1965.

Dunnette, M. D. Personnel selection and placement. Belmont, California.: Wadsworth, 1966.

Ebel, R. L. Essentials of educational measurement (2nd ed.). Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Educational Testing Service. Multiple-choice questions: A close look. Princeton, N. J.: Author, 1973.

Federal executive agency guidelines on employee selection procedures. Federal Register, 1976, 41, 51733-51759.

Fiks, A. I. Content validity for entry level testing: A progress report. Paper presented at the Content Validity II Conference, Bowling Green State University, July 1975.

Guilford, J. P. & Fruchter, B. Fundamental statistics in psychology and education (5th ed.). New York: McGraw-Hill, 1973.

Guion, R. M. Personnel testing. New York: McGraw-Hill, 1965.

Guion, R. M. Open a new window: Validities and values in psychological measurement. American Psychologist, 1974, 29, 287-296.

Guion, R. M. Content validity — the source of my discontent. Applied Psychological Measurement, 1977, 1, 1-10.

Helmstadter, G. C. Principles of psychological measurement. Englewood Cliffs, N. J.: Prentice-Hall, 1964.

Bibliography

Jensen, O. A. Improving multiple-choice test items--A manual for examiners (Personnel Report No. 603). Chicago: Public Personnel Association, 1960.

Lawshe, C. H. & Balma, M. J. Principles of personnel testing. New York: McGraw-Hill, 1966.

Lennon, R. T. Assumptions underlying the use of content validity. Educational and Psychological Measurement, 1956, 16, 294-304.

Mussio, S. J. & Smith, M. K. Content validity: a procedural manual. Chicago: International Personnel Management Association, 1973.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Payne, S. P. Guide to the development of written job-knowledge promotion tests for firefighters (Technical Memorandum 76-16). Washington, D. C.: U. S. Civil Service Commission, Personnel Research and Development Center, 1976.

Plumlee, L. B. Item writing guide, sample items, and examples of item modification. Washington, D. C.: U. S. Civil Service Commission, Personnel Research and Development Center, 1976.

Prien, Erich P. Job analytic strategies and their implications for content validity. Paper presented at the 83rd Annual Convention of the American Psychological Association, Chicago, September 1975.

Ramirez, D. A. & Lotera, R. J. Job analysis for personnel selection — a manual. Sacramento, California: California State Personnel Board, 1973.

Schwartz, D. J. Content validation. Paper presented at the International Personnel Management Association Selection Specialists' Symposium, Chicago, July 1976.

State of Wisconsin, Bureau of Personnel. Selection-oriented job analysis procedures. Department of Administration, State Bureau of Personnel, September 1972.

Tenopyr, M. L. Content-construct confusion. Paper presented at the Content Validity II Conference, Bowling Green State University, July 1975.

Thorndike, R. L. (Ed.). Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

U. S. Civil Service Commission. Job analysis for improved job-related selection (BIPP 152-63). Washington, D. C.: Author, 1975.

U. S. Civil Service Commission. Job analysis — developing and documenting data (BIPP 152-35) Washington, D. C.: U. S. Government Printing Office, 1973.

U. S. Department of Labor. Handbook for analyzing jobs. Washington, D. C.: U. S. Government Printing Office, 1972.

Wisner, R. W. The Kirkland case--its implications for personnel selection. Public Personnel Management, 1976, July-August, 263-267.

Wollack, S. Content validity: Its legal and psychometric basis. Manuscript submitted for publication, 1976.

Wright, G. H. (Ed.). Public sector employment selection. Chicago: International Personnel Management Association, 1974.

Appendix A

Sample Selection Plan Format

KSAO	✓ - screen-out X - ranking	Level	Measurement Instrument	Rationale

Method for combining scores:

rationale:

If compensatory model complete following:

Measurement Instrument	Weight	Rationale

Appendix B

Summary of Documentation File Contents

Evidence needed to support claims of job relatedness and content validity includes, at a minimum, the following records:

1. Job Analysis Materials. Document the method, instruments used, sources consulted, credentials of analysts, relevant identifying information, and results.

2. List of Knowledges, Skills, Abilities and Other Worker Characteristics. Document how linked to job, adequacy of coverage, the importance, level and type (ranking or screen-out) of each KSAO, and the criteria and method for making these determinations.

3. Selection Plan. Summarize method and adequacy of measurement of each KSAO (see sample form, Appendix A), and define each KSAO operationally.

4. Test Development Process. Document credentials of panel members, rationale for all decisions made, and log of panel sessions; include such documents as resumes, signed security forms, and signed review copies of final forms of the test content outline and assembled test.

5. Test Content Outline. Specify the correspondence between the job content domain and the test content domain.

6. Test Item Cards. While these need not be retained in the documentation file, these cards serve to document the technical adequacy of the items, their adherence to the specifications of the test plan and the accuracy of the answer key, and are a necessary part of the documentation of the test.

7. Copy of Final Version of Written Test. Include answer key and a copy of all related documents such as directions for conducting the test, sample questions, and scoring instructions (including rationale for setting cut scores).

8. Statistical Analyses of Test Scores. May include records of mean, standard deviation and frequency distribution of scores, item analysis data, and reliability studies. (In the absence of such records, the original answer sheets should be retained in the documentation file.)

9. Review Records. Document periodic review of test (if applicable).

Appendix C

Sample Security Form

Security of Civil Service Examination Material

The Personnel Research and Development Center of the Civil Service Commission requires your cooperation in preserving the confidential nature of all examination materials you review.

The following controls are suggested:

1. Submit to the Commission the questions that you revise or construct, and retain no copy or notes pertaining to them or to any written sources used. (All test materials should be returned.)

2. Make no personal or file copies of questions you are asked to review or of your comments on them.

3. Keep all materials pertaining to examination questions under lock and key when they are not in actual use.

4. Permit no other person to have access to examination materials without prior clearance with the Personnel Research and Development Center.

5. Keep in mind that one may have a tendency to inadvertently give an undue advantage to persons supervised, taught, or merely talked to and who may take examinations containing some of the questions you have worked on.

Because of the importance of security considerations in examinations for the public service, we should appreciate your signing this request form and returning it ti us as indication that the foregoing types of safeguards, and any additional ones that seem desirable, will be applied. We should also appreciate your notifying us if in any case circumstances beyond your control cause you to have any doubts as to whether the confidential nature of the materials has been maintained.

Signature: _____ Date: _____

Agency: _____

Appendix D

Sample Test Content Outline

Type of Test	:	Job Knowledge
Target Job	:	Accountant I
Test Objective	:	to measure applicants' present knowledge of accounting terms and procedures. Test scores will be used to rank applicants in the last phase of a multiple hurdle selection program which also includes assessment of experience and performance on a work sample of accounting problems.
Item Format	:	Multiple-choice (4 choice)
Test Length	:	100 items
Time allowed	:	2 hours

Sampling Plan

Level of Measurement: demonstration of knowledge at the level needed for the successful performance of a diversity of routine accounting tasks in a municipal government environment under close supervision.

Knowledges sampled :		No. of items
Area I.	Balance Sheet	8
	1. Classification of balance sheet items	4
	2. Relation between balance sheet and income statement	2
	3. Debit and credit balance	2
Area II.	Cost Accounting	6
	1. Cost analysis	3
	2. Billing and collection procedures	2
	3. Inventory pricing procedures (sample only	1
	4. Depreciation allowance one)	
Area III.	Assets	6
	1. Fixed assets: accounting treatment of purchase and disposition	4
	2. Intangible assets: Accounting treatment of goodwill, leasehold improvements, patents and trademarks	2
.	.	.
.	.	.
.	.	.
.	.	.
Area X.	Ratios	2
	1. Types: working capital ratio; acid test ratio	1
	2. Capital turnover and inventory turnover	1